Discussion of "On simulation and properties of the stable law" by L. Devroye and L. James

Stat Methods Appl (2014) 23:365–369
DOI 10.1007/s10260-014-0270-y

Discussion of "On simulation and properties of the stable law" by L. Devroye and L. James

Stefano Favaro · Bernardo Nipoti

Accepted: 16 May 2014 / Published online: 10 June 2014

Abstract We contribute to the discussion of the paper by Devroye and James, by reviewing some of the most meaningful results that relate the unilateral stable distribution with the asymptotic behavior of the so-called Ewens-Pitman sampling model. Our focus is then on how these results have been exploited in the context of Bayesian nonparametric inference for species sampling problems.

Keywords Tilted stable distribution · Ewens-Pitman sampling model · Bayesian nonparametric inference · Two-parameter Poisson-Dirichlet process · Asymptotic credible intervals

We congratulate Luc Devroye and Lancelot James for their comprehensive and stimulating paper on stable laws. The paper provides a useful survey on properties and distributional results for the stable law, as well as for related distributions, and exact sampling algorithms. We aim at contributing to the discussion of the paper by Devroye and James with a review of some of the most meaningful results that relate the unilateral stable distribution with the asymptotic behaviour of the so-called Ewens-Pitman sampling model introduced by Pitman (1995). These results have recently found several applications in the context of Bayesian nonparametric inference for species sampling problems arising from ecology, biology, genetic, linguistic, etc. In such a context exact sampling algorithms for polynomially tilted positive stable distribution provide a useful tool in order to concretely implement some inferential procedures.

S. Favaro (✉) · B. Nipoti
University of Torino and Collegio Carlo Alberto, C.so Unione Sovietica 218/bis, 10134 Torino, Italy
e-mail: stefano.favaro@unito.it

B. Nipoti
e-mail: bernardo.nipoti@unito.it

1 Asymptotics for the Ewens–Pitman sampling model

Among various definitions of the Ewens–Pitman sampling model, a simple and intuitive one was introduced by Zabell (1997) in terms of the following urn scheme. Let $\alpha \in (0, 1)$ and consider an urn that initially contains a black ball with mass $\theta > 0$. Balls are drawn from the urn successively with probabilities proportional to their masses. When a black ball is drawn, it is returned to the urn together with a black ball of mass α and a ball of a new color with mass $(1-\alpha)$. The color of the new ball is sampled from a nonatomic probability measure ν. When a non-black ball is drawn, it is returned to the urn with an additional ball of the same color with mass one. Let $(X_i)_{i\geq 1}$ be the sequence of non-black colors, then

$$\mathbb{P}[X_{n+1} \in \cdot \mid X_1, \ldots, X_n] = \frac{\theta + j\alpha}{\theta + n}\nu(\cdot) + \frac{1}{\theta + n}\sum_{i=1}^{j}(n_i - \alpha)\delta_{X_i^*}(\cdot), \qquad (1)$$

where $(X_1^*, \ldots, X_j^*)$ are the j distinct colors with frequencies $\mathbf{n} = (n_1, \ldots, n_j)$. The predictive distribution (1) was introduced in Pitman (1995) for any $\alpha \in (0, 1)$ and $\theta > -\alpha$, and it is referred to as the Ewens–Pitman sampling model. In particular, Pitman (1995) showed that the sequence $(X_i)_{i\geq 1}$ generated by (1) is exchangeable and its de Finetti measure corresponds to the distribution of the two parameter Poisson–Dirichlet process $\tilde{P}_{\alpha,\theta}$ in Perman et al. (1992). Formally,

$$\begin{aligned} X_i \mid \tilde{P}_{\alpha,\theta} &\overset{\text{iid}}{\sim} \tilde{P}_{\alpha,\theta} \qquad i = 1, \ldots, n \\ \tilde{P}_{\alpha,\theta} &\sim \Pi, \end{aligned} \qquad (2)$$

for any $n \geq 1$. As $\alpha \to 0$ the urn model generating the X_i's reduces to the one introduced by Hoppe (1984), and the Ewens–Pitman sampling model reduces to the celebrated sampling model by Ewens (1972). Accordingly, as $\alpha \to 0$ the two parameter Poisson–Dirichlet process reduces to the Dirichlet process by Ferguson (1973). The Ewens–Pitman sampling model plays an important role in various research areas such as population genetics, machine learning, Bayesian nonparametrics, combinatorics and statistical physics. See the monograph by Pitman (2006) and references therein for a comprehensive account on these sampling models.

According to (1) and (2) a sample $(X_1, \ldots, X_n)$ from $\tilde{P}_{\alpha,\theta}$ induces a random partition of the set $\{1, \ldots, n\}$ into K_n blocks with frequencies $\mathbf{N}_n = (N_1, \ldots, N_{K_n})$. See Pitman (1995) for details. The unilateral stable distribution arises in the large n asymptotic behaviour of K_n. Specifically, for any $\alpha \in (0, 1)$, let f_α be the density function of the unilateral stable random variable, and consider a random variable $Z_{\alpha,q}$, for any real $q > -1$, with density function

$$f_{Z_{\alpha,q}}(z) = \frac{\Gamma(q\alpha + 1)}{\alpha\Gamma(q + 1)} z^{q-1-1/\alpha} f_\alpha(z^{-1/\alpha}). \qquad (3)$$

The random variable $Z_{\alpha,q}^{-1/\alpha}$ is the so-called polynomially tilted unilateral α-stable random variable. For any $\alpha \in (0,1)$ and $\theta > -\alpha$, Pitman (1996) showed that, as $n \to +\infty$,

$$\frac{K_n}{n^\alpha} \xrightarrow{\text{a.s.}} Z_{\alpha,\theta/\alpha}. \tag{4}$$

Furthermore, let $M_n(l)$ be the number of blocks with frequency $l \geq 1$ such that $K_n = \sum_{1 \leq l \leq n} M_n(l)$ and $n = \sum_{1 \leq l \leq n} l M_n(l)$. Then Pitman (2006) showed that, as $n \to +\infty$,

$$\frac{M_n(l)}{n^\alpha} \xrightarrow{\text{a.s.}} \frac{\alpha(1-\alpha)_{(l-1)}}{l!} Z_{\alpha,\theta/\alpha}, \tag{5}$$

where $(a)_{(n)} := a(a+1)\cdots(a+n-1)$. Weak convergence versions of (4) and (5) can also be derived from general asymptotic results for urn model with weighted balls. See Proposition 16 in Flajolet et al. (2006) and Theorem 5 in Janson (2006) for details. The fluctuation limits (4) and (5) display the crucial role of the parameter α in determining both the clustering structure and the large n asymptotic behaviour of K_n: the bigger α the flatter is the distribution of K_n.

2 On species sampling problems in Bayesian nonparametrics

From a Bayesian perspective, the hierarchical framework (2) provides a nonparametric model for the individuals X_i's of a population with infinite species X_i^*, where Π is the prior distribution on the species compositions. Under this framework, a novel Bayesian nonparametric approach for making inference on species sampling problems was introduced in Lijoi et al. (2007). Such an approach consists in evaluating, conditionally on the random partition $(K_n, \mathbf{N}_n)$ induced by an initial sample $(X_1, \ldots, X_n)$ from $\tilde{P}_{\alpha,\theta}$, the distribution of statistics of interest from an additional unobserved sample $(X_{n+1}, \ldots, X_{n+m})$. Lijoi et al. (2007) focussed on the conditional distribution of the number $K_m^{(n)}$ of new species in $(X_{n+1}, \ldots, X_{n+m})$, whereas Favaro et al. (2013) considered the problem of determining the conditional distribution of the number $M_m^{(n)}(l)$ of species with frequency $l \geq 1$ in $(X_1, \ldots, X_{n+m})$. Expected values of these conditional, or posterior, distributions take on the interpretation of the Bayesian nonparametric estimators of the number of new distinct species and the number of district species with frequency l generated by the additional sample. Throughout this section we write $X \mid Y$ to denote, with a slight abuse of notation, a random variable whose distribution is the conditional distribution of X given Y.

Let $(X_1, \ldots, X_n)$ be a sample from $\tilde{P}_{\alpha,\theta}$ featuring $K_n = j$ species with corresponding frequencies $\mathbf{N}_n = \mathbf{n}$. The large m asymptotic behaviour of $K_m^{(n)} \mid (K_n = j, \mathbf{N}_n = \mathbf{n})$ is studied in Favaro et al. (2009). Specifically, for any $j \leq n$, we introduce the random variable $Z_{\alpha,\theta,j}^{(n)} := B_{j+\theta/\alpha, n/\alpha - j} Z_{\alpha,(\theta+n)/\alpha}$ where $B_{a,b}$ is a Beta random variable with parameter (a,b) and $Z_{\alpha,q}$ is a random variable, independent of $B_{a,b}$, and distributed as in (3). Then, as $m \to +\infty$,

$$\frac{K_m^{(n)}}{m^\alpha} \mid (K_n = j, \mathbf{N}_n = \mathbf{n}) \xrightarrow{\text{a.s.}} Z_{\alpha,\theta,j}^{(n)}. \tag{6}$$

The large m asymptotic behaviour of $M_m^{(n)}(l)$, conditionally on the random partition $(K_n, \mathbf{N}_n)$, follows from (6) and Corollary 21 in Gnedin et al. (2007). Specifically, as $m \to +\infty$,

$$\frac{M_m^{(n)}(l)}{m^\alpha} \mid (K_n = j, \mathbf{N}_n = \mathbf{n}) \xrightarrow{\text{a.s.}} \frac{\alpha(1-\alpha)_{(l-1)}}{l!} Z_{\alpha,\theta,j}^{(n)}. \tag{7}$$

The fluctuation limits (6) and (7) take on the interpretation of the posterior counterpart of (4) and (5), respectively. See Favaro and Feng (2014) for a generalization of (6) to the total number of species generated by the additional sample, namely the number $K_m^{(n)}$ of new species plus the number of species which coincide with species already detected in the initial observed sample. See also Theorem 4 in Favaro et al. (2013) for weak convergence versions of (7).

The fluctuation limits (6) and (7) provide a useful tool in order to approximate, for large m, the posterior distributions of $K_m^{(n)}$ and $M_m^{(n)}(l)$. Indeed, as pointed out in Favaro et al. (2009) and Favaro et al. (2013), there are situations of practical interest where j, n and m are very large and the computational burden for evaluating these posterior distributions becomes overwhelming. As an example, Favaro et al. (2009) and Cesari et al. (2012) applied (6) and (7) in order to obtain asymptotic credible intervals for posterior estimators of $K_m^{(n)}$ and $M_m^{(n)}(l)$, respectively. Their approach consisted in evaluating, via simulation, appropriate quantiles of the limiting posterior distributions in order to obtain an approximate evaluation of the credible intervals. Of course such a procedure involves sampling from a random variable with density function (3). To this end, one can either resort directly to the exact algorithm for sampling from polynomially tilted positive stable distributions in Devroye (2009) or alternatively, by means of a gamma-type augmentation, on can exploit the exact algorithms for exponentially tilted positive stable distributions in Devroye (2009) and Hofert (2011).

3 Concluding remarks

While our discussion focussed on the Ewens–Pitman sampling model, the unilateral stable distribution also arises in the study of the asymptotic behavior of a more general class of sampling models, the so-called Gibbs-type sampling models introduced by Gnedin and Pitman (2006). This class generalizes the Ewens–Pitman sampling model as follows. Let $\alpha \in (0, 1)$ and let $V = (V_{n,j})_{j \le n, n \ge 1}$ be nonnegative weights satisfying $V_{n,j} = V_{n+1,j+1} + (n - j\alpha) V_{n+1,j}$, with $V_{1,1} = 1$. Then, a Gibbs-type sampling model with parameter (α, V) is defined as

$$\mathbb{P}[X_{n+1} \in \cdot \mid X_1, \ldots, X_n] = \frac{V_{n+1,j+1}}{V_{n,j}} \nu(\cdot) + \frac{V_{n+1,j}}{V_{n,j}} \sum_{i=1}^{j} (n_i - \alpha) \delta_{X_i^*}(\cdot) \tag{8}$$

for any $n \ge 1$, with $X_1^*, \ldots, X_j^*$ being the j distinct observations in $(X_1, \ldots, X_i)$ with frequencies $(n_1, \ldots, n_j)$. If $V_{n,j} = \prod_{0 \le i \le j-1} (\theta + i\alpha)/(\theta)_{(n)}$, for any $\theta > -\alpha$, then

(8) reduces to (1). Pitman (2003), and more recently James (2013), provided details for generalizing (4) and (6) to the framework of Gibbs-type species sampling models. Accordingly, the corresponding generalizations of (5) and (7) can be easily derived by a direct application of Corollary 21 in Gnedin et al. (2007).

Acknowledgments Stefano Favaro is supported by the European Research Council (ERC) through StG N-BNP 306406.

References

Cesari O, Favaro S, Nipoti B (2012) Posterior analysis of rare variants in Gibbs-type species sampling models. Preprint

Devroye L (2009) Random variate generation for exponentially and polynomially tilted stable distributions. ACM Trans Model Comput Simul. doi:10.1145/1596519.1596523

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87–112

Favaro S, Lijoi A, Mena RH, Prünster I (2009) Bayesian nonparametric inference for species variety with a two parameter Poisson–Dirichlet process prior. J R Stat Soc Ser B 71:993–1008

Favaro S, Lijoi A, Prünster I (2013) Conditional formulae for Gibbs-type exchangeable random partitions. Ann Appl Probab 23:1721–1754

Favaro S, Feng S (2014) Asymptotics for the number of blocks in a conditional Ewens–Pitman sampling model. Electron J Probab 19:1–15

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230

Flajolet P, Dumas P, Puyhaubert V (2006) Some exactly solvable models of urn process theory. In: Discrete Math. Theor. Comput. Sci. Proceedings of the fourth colloquium on mathematics and computer science, pp 59–118

Gnedin A, Pitman J (2006) Exchangeable Gibbs partitions and Stirling triangles. J Math Sci 138:5674–5685

Gnedin S, Hansen B, Pitman J (2007) Notes on the occupancy problem with infinitely many boxes: general asymptotics and power law. Probab Surv 4:146–171

Hofert M (2011) Efficiently sampling nested Archimedean copulas. Comput Stat Data Anal 55:57–70

Hoppe FM (1984) Pólya-like urns and the Ewens sampling formula. J Math Biol 20:91–94

James LF (2013) Stick-breaking PG(α, ζ)-generalized gamma processes. Preprint arXiv:1308.6570

Janson S (2006) Limit theorems for triangular urn schemes. Probab Theory Rel Fields 134:417–452

Lijoi A, Mena RH, Prünster I (2007) Bayesian nonparametric estimation of the probability of discovering new species. Biometrika 94:769–786

Perman M, Pitman J, Yor M (1992) Size-biased sampling of Poisson point processes and excursions. Probab Theory Rel Fields 92:21–39

Pitman J (1995) Exchangeable and partially exchangeable random partitions. Probab Theory Rel Fields 102:145–158

Pitman J (1996) Partition structures derived from Brownian motion and stable subordinators. Bernoulli 3:79–66

Pitman J (2003) Poisson–Kingman partitions. In: Goldstein DR (ed. Science and statistics: a Festschrift for terry speed, Lecture Notes Monograph Series, vol 40, pp 1–34. IMS, Beachwood, OH

Pitman J (2006) Combinatorial stochastic processes. Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875. Springer, New York

Zabell SL (1997) The continuum of inductive methods revisited. In: The cosmos of science, Pittsburgh-Konstanz Series in the philosophy and history of science, pp 351–385